ChinaXiv合作期刊

心理学报 2022, Vol. 54, No. 11, 1403−1415
Acta Psychologica Sinica
© 2022 中国心理学会
https://doi.org/10.3724/SP.J.1041.2022.01403

多级属性 *Q* 矩阵的验证与估计*

秦春影[1,2] 喻晓锋[1]

([1]江西师范大学心理学院, 南昌 330022) ([2]南昌师范学院数学与信息科学学院, 南昌 330032)

摘 要 多级属性是将诊断测验中传统的二值(即两种水平, 通常定义为 0 和 1)属性定义为多值(多个水平可以为 0, 1, …), 它不但可以描述学生对于知识属性是否掌握, 而且可以描述学生在属性上的掌握程度, 这样使得诊断测验能提供给被试更丰富的知识掌握详情。本文将适用于二级属性 ***Q*** 矩阵的统计量(*S* 统计量)拓展到多级属性下的 ***Q*** 矩阵验证和估计, 在两种常见的条件下, 设计了两种估计算法：联合估计算法和在线估计算法。模拟实验结果表明：联合估计算法适用于对专家界定的初始 ***Q*** 矩阵进行验证, 当初始 ***Q*** 矩阵中包含较少的错误时, 通过联合估计算法有很大可能恢复正确的 ***Q*** 矩阵; 在线估计算法适用于对"新项目"进行属性向量和项目参数的在线标定, 基于一定数量的"基础项目", 在线估计算法对于新项目的估计也能达到较满意的成功率。实证数据分析则进一步展示了该方法的使用。

关键词 多级属性, ***Q*** 矩阵, p-DINA 模型, *S* 统计量
分类号 B841

1 引言

随着社会的发展, 教育和心理测验已经不满足于单一的总体评价(overall assessment)。认知诊断评价(cognitive diagnosis assessment, CDA)可以提供学生在知识上的掌握详情, 已受到社会的广泛关注(Leighton & Gierl, 2007; Tatsuoka, 2009; Rupp et al., 2010; 罗照盛, 2019; von Davier & Lee, 2019)。传统的测验, 如基于经典测验理论(classical test theory, CTT)或基于项目反应理论(item response theory, IRT)的测验都仅仅提供学生的总体分数或能力, 除了这个总体评价之外, CDA 还可以提供学生的知识状态(knowledge state, KS), 这个知识掌握状态可以对学生的学习、教师的教学和教学效果的评价起到很好的指导和参考作用。

通常情况下, CDA 中学生对知识的掌握情况是用 0 或 1 来描述, 1 表示学生掌握了某个知识, 0 表示没有掌握, 即学生对知识的掌握仅仅有 2 个水平。文献中通常把 CDA 中细粒度的知识用属性(attribute; Leighton et al., 2004)来描述, 学生在这多个属性上的掌握情况就是学生的 KS。因此, 学生的 KS 通常是一个二值向量。将学生对属性的掌握情况用 0 和 1 来描述的好处是相对简单, 容易解释, 但是却也相对粗糙, 不能准确刻画学生在属性上的掌握程度, 因为两个在某属性上的状态都为 0 的学生之间还是有掌握程度上的区别的。也正是因为如此, 有很多研究者考虑将属性的二种取值考虑设置成多种取值(Karelitz, 2004; von Davier, 2008; Chen & de la Torre, 2013; Sun et al., 2013; 蔡艳, 涂冬波, 2015; 涂冬波, 蔡艳, 2015; 詹沛达 等, 2016; Zhan et al., 2020; Shang et al., 2021)。实际应用中, 有很多情况都是对知识属性的多水平要求和考查, 比如《全日制义务教育数学课程标准(修改稿)》中就使用了"了解(认识)"、"理解"、"掌握"和"运用"这 4 个顺序类别词汇来表述知识技能目标的不同水平。因此, 多级属性能够对学生做出更为精细地划分,

收稿日期: 2021-10-06
* 全国教育科学规划项目(BGA210060); 教育部教育考试院"十四五"规划支撑专项课题(NEEA2021050);江西省社会科学基金项目(21JY06); 江西省高校人文社会科学项目(XL20202); 南昌市教育大数据智能技术重点实验室(2020-NCZDSY-012); 江西省教育厅科技项目(GJJ212602, GJJ191691, GJJ191128)资助。
通信作者: 喻晓锋, E-mail: xyu6@jxnu.edu.cn

chinaXiv:202303.08463v1

ChinaXiv合作期刊

将属性定义成多级的诊断测验具有现实应用价值和前景。

也正是因为如此，研究者们对基于多级属性的CDA展开了研究，有针对性地开发了诊断模型，比如 Karelitz (2004)构建了基于顺序类别属性编码(ordered-category attribute coding, OCAC)的诊断模型 OCAC-DINA，并且对 $\boldsymbol{Q}$ 矩阵中存在缺失时的参数估计和分类进行研究；还有基于其它诊断模型所开发的多级属性模型，像 RRUM 下的多级属性模型(Templin, 2004)，LCDM 下的多级属性模型(Templin & Bradshaw, 2014)；GDM 下的多级属性模型(Haberman et al., 2008; von Davier, 2008)；Zhan 等人(2020)构建了高阶的多级属性的诊断模型等；与前面这些研究不同的是，Shang 等人(2021)借鉴多维 IRT 的思想，定义连续的多级属性，并且构建了可以处理连续多级属性的诊断模型。同传统的CDA 一样，多级属性 CDA 中的 $\boldsymbol{Q}$ 矩阵的作用也十分关键，它的正确性会直接影响模型参数的识别、被试的分类乃至整个测验的信度和效度。并且更重要的是，在实际应用中，仅仅由专家界定的 $\boldsymbol{Q}$ 矩阵很容易出现错误或专家意见不一致的情况(de la Torre, 2008; DeCarlo, 2012; Liu et al., 2012; 喻晓锋等, 2015a; Yu & Cheng, 2020)。从目前已有的研究来看，研究者们采用的多级属性 $\boldsymbol{Q}$ 矩阵大都是由专家界定或模拟生成，通常假定它是正确的，没有对它的正确性或合适性进行验证，还缺乏对多级属性 $\boldsymbol{Q}$ 矩阵的验证和估计方法进行研究。因此，迫切需要研究客观的方法来对其正确性进行验证或估计。本研究拟将适合二级属性下 $\boldsymbol{Q}$ 矩阵的验证和估计方法拓展到适合多级属性 $\boldsymbol{Q}$ 矩阵的情况，研究客观的验证或估计多级属性 $\boldsymbol{Q}$ 矩阵的方法，以期能促进多级属性 CDA 的发展。

2 多级属性 Q 矩阵及诊断模型

在正式介绍多级属性 $\boldsymbol{Q}$ 矩阵的估计算法之前，首先对多级属性 $\boldsymbol{Q}$ 矩阵及对应的诊断模型进行介绍。

2.1 多级属性 Q 矩阵

为方便介绍，在不引起误解的情况下，将仅仅有 0, 1 两种取值的属性称为二级属性(binary attribute)，仅仅由二值属性构成的 $\boldsymbol{Q}$ 矩阵称为二级属性 $\boldsymbol{Q}$ 矩阵(binary-attribute $\boldsymbol{Q}$ matrix, BQM)，用 $\boldsymbol{Q}_B$ 表示，将采用 $\boldsymbol{Q}_B$ 的 CDA 记为 BCDA；将可取 0, 1, 2, …多种值的属性称为多级属性，包含多级属性的 $\boldsymbol{Q}$ 矩阵称为多级属性 $\boldsymbol{Q}$ 矩阵(polytomous-attribute $\boldsymbol{Q}$ matrix, PQM)，用 $\boldsymbol{Q}_P$ 表示，将采用 $\boldsymbol{Q}_P$ 的 CDA 记为 PCDA。$\boldsymbol{Q}_P$ 是一个 $J\times K$ 的矩阵，其中 J 和 K 分别表示项目数和属性数，$\boldsymbol{Q}_P$ 中的元素记为 q_{jk}，与二级的 $\boldsymbol{Q}_B$ 不同，$\boldsymbol{Q}_P$ 中的 q_{jk} 有 M_k 个水平，取值空间为 $0,1,\cdots M_k-1$。

下面以一简单的多级属性 $\boldsymbol{Q}$ 矩阵(Karelitz, 2004)为例，这个 $\boldsymbol{Q}_P$ 中有 4 个项目，共考察了 2 个属性，其中属性 1 和属性 2 皆有 0, 1, 2, 3, 4 共 5 个水平。

$$\begin{array}{c} \\ \boldsymbol{Q}_P = \end{array}\begin{array}{c} \\ \text{项目1} \\ \text{项目2} \\ \text{项目3} \\ \text{项目4} \end{array}\begin{array}{c} \begin{array}{cc}\text{属性1} & \text{属性2}\end{array} \\ \begin{bmatrix} 0 & 4 \\ 2 & 3 \\ 3 & 2 \\ 4 & 1 \end{bmatrix} \end{array} \tag{1}$$

如果属性按按传统的二级方式，用 0 作为截断点来对属性进行划分，则其所对应的 $\boldsymbol{Q}$ 矩阵如(2)所示。

$$\begin{array}{c} \\ \boldsymbol{Q}_B = \end{array}\begin{array}{c} \\ \text{项目1} \\ \text{项目2} \\ \text{项目3} \\ \text{项目4} \end{array}\begin{array}{c} \begin{array}{cc}\text{属性1} & \text{属性2}\end{array} \\ \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \end{array} \tag{2}$$

当测验考虑 K 个属性，若采用二级属性，最多能将学生总体分为 2^K 类，而采用多级属性(各属性可能的取值个数记为 M_k)，则最多可将学生总体分为 $\prod_{k=1}^{K} M_k$ 类，很明显 $\prod_{k=1}^{K} M_k$ 总是大于 2^K 的。举个简单的例子，假设测验考察 2 个属性，如果采用二级属性，最多可以将学生分为 $2^2=4$ 类，如果采用 5 值属性，则可以将学生分为 $5^2=25$ 类。

2.2 多级属性下的诊断模型

已开发的适合多级属性的诊断模型主要有OCAC-DINA (Karelitz, 2004), LCDM 下的多级属性模型(Templin & Bradshaw, 2014), GDM 对应的多级属性诊断模型(Haberman et al., 2008; von Davier, 2008)，基于 G-DINA 框架下的多级属性模型，比如 Chen 和 de la Torre (2013)，蔡艳和涂冬波(2015)，高阶的多级属性模型(Zhan et al., 2020)，连续的多级属性诊断模型(Shang et al., 2021)等。在这里，为节省篇幅，仅仅介绍与本文有关的 pG-DINA 和 p-DINA 模型。

pG-DINA (polytomous generalized deterministic inputs, noisy, “and” gate)即 G-DINA 模型的多级属性版本(Chen & de la Torre, 2013)。为方便介绍并且

chinaXiv:202303.08463v1

ChinaXiv合作期刊

不失一般性，假定测验中所有属性有相同的水平数，即 $M_k = M$，相关的符号与Chen 和de la Torre (2013), de la Torre (2011)保持一致。其中 $K_j^* = \sum_{K=1}^{K} I(q_{jk} > 0)$ 用来表示项目 j 所考察的属性的个数，为方便介绍，假设项目 j 考察的属性恰好是前 K_j^* 个属性，项目 j 所需要的属性可以表示为简化的向量 $\alpha_{lj}^* = (\alpha_{l1},\cdots,\alpha_{lK_j^*})'$，其中 $l = 1,\cdots,M^{K_j^*}$，α_{lj}^* 中的元素的取值范围是 $[0, M-1]$，这样一来，项目 j 需要考虑的属性向量个数由 M^K 下降到 $M^{K_j^*}$，即将那些没有考察的属性不予考虑，当然这样的简化也可以提高参数估计的速度。

在 p-DINA 模型下，每个项目都将学生分为两类，即掌握项目的学生(掌握了题目所考察的属性，并且考生对属性的掌握水平都不低于题目所考察的水平)和未掌握项目的学生(没有完全掌握题目所考察的属性，或者考生对属性的掌握至少有一个低于题目所考察的水平)。进一步，对于项目 j 来说，若 $q_{jk} = m$，则学生在该属性上的掌握情况 α_{lk} 可以压缩为一个二级的状态，即

$$\alpha_{lk}^{**} = \begin{cases} 0 & if\ \alpha_{lk} < q_{jk} \\ 1 & otherwise \end{cases} \quad (3)$$

记 $\alpha_{lj}^{**} = (\alpha_{j1}^{**},\cdots,\alpha_{lK_j^*}^{**})$ 为压缩后的属性掌握向量，其中 $l = 1,\cdots,2^{K_j^*}$，这样就将被试参数的个数由 $M^{K_j^*}$ 下降到 $2^{K_j^*}$，关于这部分的详细过程请参考 Chen 和 de la Torre (2013)的 Table 2。

在 pG-DINA 模型的饱和形式下，属性向量为 α_{lj}^{**} 的被试正确作答项目 j 的概率为

$$P\left(X_j = 1 | \alpha_{lj}^{**}\right) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk}^{**} + \sum_{k'>k}^{K_j^*}\sum_{k=1}^{K_j^*} \delta_{jkk'}\alpha_{lk}^{**}\alpha_{lk'}^{**} + \cdots + \delta_{j1,\cdots K_j^*}\prod_{k=1}^{K_j^*}\alpha_{lK_j^*}^{**}, \quad (4)$$

其中 δ_{j0} 表示项目 j 的截距参数，即学生未掌握该项目的任何属性时的作答概率；δ_{jk} 是属性 k 的主效应，即学生掌握属性 k 所带来的正确作答概率增加的部分；$\delta_{jkk'}$ 是属性 k 和 k' 的交互效应，即学生同时掌握属性 k 和 k' 所带来的正确作答概率增加的部分；$\delta_{j1,\cdots K_j^*}$ 是属性 $1,\cdots,K_j^*$ 的交互效应。

当只考虑截距和 K_j^* 阶交互效应时，则 pG-DINA 就变成了 p-DINA 模型；当只考虑截距和 K_j^* 个属性的主效应时，则则 pG-DINA 就变成了 pA-CDM 模型；其它的模型，如 p-DINO, pR-RUM 等模型也可以在 pG-DINA 模型上增加约束获得。因为 p-DINA 模型相对简单，所以本研究中基于 p-DINA 模型研究多级属性 $\boldsymbol{Q}$ 矩阵的估计和验证。

3　多级属性 Q 矩阵的估计方法

在介绍多级属性 $\boldsymbol{Q}$ 矩阵的估计之前，首先对二级属性 $\boldsymbol{Q}$ 矩阵的估计方法作个简单回顾。在 BCDA 中，有很多研究者对 $\boldsymbol{Q}$ 矩阵的验证和估计进行了深入的研究，比如 de la Torre (2008)提出的 δ 方法，涂冬波等人(2012)采用的 γ 方法，DeCarlo (2012)采用的贝叶斯方法, Liu 等人(2012)提出的基于 S 统计量的方法，Xiang (2013)采用的惩罚估计进行探索的方法, Chung (2014)采用 MCMC 估计进行探索的方法，喻晓锋等人(2015b)采用的基于 S 统计量的联合估计方法; de la Torre 和 Chiu (2016)基于 G-DINA 模型提出了一种经验的 $\boldsymbol{Q}$ 矩阵验证的方法；Wang 等人(2020)在已知 $\boldsymbol{Q}$ 矩阵中部分题目的属性定义基础上，对几种基于似然比检验的方法进行了比较；Yu 和 Cheng (2020)考虑了一种基于残差的 $\boldsymbol{Q}$ 矩阵验证方法等。

在众多的方法中，基于 S 统计量的方法是完全基于作答数据的客观方法，并且 Liu 等人对它相应的理论基础进行了严格的证明(Liu et al., 2013; Xu, 2013)，该方法不依赖于具体的诊断模型和测验计分方式，有非常好的拓广性。因此本研究拟将 S 统计量拓展到适合多级属性 $\boldsymbol{Q}$ 矩阵的估计。考虑实际应用中可能出现的两种情况，第一种情况与 Liu 等人(2012)相同，即已经由专家界定了 $\boldsymbol{Q}$ 矩阵，记为 $\boldsymbol{Q}_0$，只是还不确定 $\boldsymbol{Q}_0$ 是否完全正确(完全正确是指 $\boldsymbol{Q}_0$ 中每个项目的属性向量都正确)，因此需要采用客观的方法进行验证，这里拟对 S 统计量进行拓广，使之适合多级属性情况下的 $\boldsymbol{Q}$ 矩阵验证；考虑的另一种情况是手头只有少数的项目属性向量已经界定，有更多的“新项目”需要属性向量的定义，即考虑多级属性情况下的 $\boldsymbol{Q}$ 矩阵估计。为方便介绍，记适合二级属性的 S 统计量为 S_B，适合多级属性 $\boldsymbol{Q}$ 矩阵的 S 统计量为 S_P。

下面介绍基于 S_P 统计量的多级属性 $\boldsymbol{Q}$ 矩阵的估计。

3.1　基于 S_P 统计量的多级属性 Q 矩阵估计

构建 S 统计量的核心是 T 矩阵，T 矩阵中的元素描述的是不同能力的考生在测验单个题目上或所有可能题目组合上的期望正确作答概率，它将期

chinaXiv:202303.08463v1

ChinaXiv合作期刊

望作答分布和模型结构联系起来了，是 $\boldsymbol{Q}$ 矩阵定义的反映，它建立起了属性分布和作答分布间的线性依赖关系(Liu et al., 2012, 2013; Qin et al., 2015)。

测验考察了 K 个属性，每个属性有 M 个水平，因此，学生的属性掌握状态有 M^K 种可能。T 矩阵一共有 M^K 列，T 矩阵的行分别对应了各类学生对单个项目、2 个项目的组合、…，J 个项目组合的正确作答概率，此时构建的 T 矩阵如下(4)所示。

$$T_p = \begin{matrix} \{1\} \\ \vdots \\ \{j\} \\ \{1,2\} \\ \vdots \\ \{1,2,\cdots,J\} \end{matrix} \begin{bmatrix} P_{\{1\},1} & P_{\{1\},2} & \cdots & P_{\{1\},M^K} \\ \vdots & \vdots & \vdots & \vdots \\ P_{\{j\},1} & P_{\{j\},2} & \cdots & P_{\{j\},M^K} \\ P_{\{1,2\},1} & P_{\{1,2\},2} & \cdots & P_{\{1,2\},M^K} \\ \vdots & \vdots & \cdots & \vdots \\ P_{\{1,2,\cdots,J\},1} & P_{\{1,2,\cdots,J\},2} & \cdots & P_{\{1,2,\cdots,J\},M^K} \end{bmatrix} \quad (4)$$

T_p 的行表示各单个项目及其所有可能的组合，共有 2^J-1 行，其中 $1 \wedge 2$ 对应的行表示同时正确作答项目 1 和项目 2 的概率；T_p 的列表示所有可能的学生类，在不考虑属性间关系的情况下，共有 M^K 列。

假设学生的总体分布记为 p，通常情况下，在没有先验信息的情况下可以把 p 按均匀分布处理，在估计过程中采用经验贝叶斯方法(de la Torre, 2009)来对其进行更新。

学生在各单个项目及其可能组合(即 $\{1\},\cdots,\{j\}$ $\{1,2\},\cdots,\{1,2,\cdots,J\}$)上的期望作答分布可以通过 $T_p \times p$ 得到，如(5)式所示，其中，$P(R^1=1|\boldsymbol{Q}',p_\alpha,\hat{\psi})$ 表示该被试总体中正确作答项目 1 的期望概率，其计算方法如(6)所示。

$$T_p \times p = \begin{pmatrix} P(R^1=1|\boldsymbol{Q}',p,\hat{\psi}) \\ \vdots \\ P(R^J=1|\boldsymbol{Q}',p,\hat{\psi}) \\ P(R^1=1,R^2=1|\boldsymbol{Q}',p,\hat{\psi}) \\ \vdots \end{pmatrix}, \quad (5)$$

$$P(R^j=1|\boldsymbol{Q}',p,\hat{\psi}) = \sum_{\alpha\in\Omega^\alpha} p_\alpha P(R^j=1|\boldsymbol{Q}',\alpha,\hat{\psi}), \quad (6)$$

其中，Ω^α 表示被试的属性掌握模式全集。另一方面，学生的观察作答分布(用 $\hat{\beta}$ 表示)可以从作答数据中得到，这里项目参数(用 $\hat{\psi}$ 表示)使用 EM 算法(de la Torre, 2011)估计，学生的知识状态(用 $\hat{\alpha}$ 表示)通过 MAP 算法(de la Torre, 2009)得到。当 $\boldsymbol{Q}$ 矩阵界定正确，各参数误差较小的情况下，根据大数定律，在被试人数足够多，即 $N\to\infty$ 时，有 $T_p^O \times p_\alpha \to \hat{\beta}$ 成立，即测验总体的观察作答分布依概率收敛于其期望分布(Liu et al., 2012, 2013; Xu, 2013)。当包含猜测和失误时，$\boldsymbol{Q}$ 矩阵中包含的错误越少，此时 $T_p^O \times p_\alpha$ 和 $\hat{\beta}$ 之间的距离应该越小，因此估计多级属性 $\boldsymbol{Q}$ 矩阵的目标函数为

$$S(\boldsymbol{Q}') = \inf \left| T_p^O(\boldsymbol{Q}')p_\alpha - \hat{\beta} \right|, \quad (7)$$

$$\hat{\boldsymbol{Q}} = \underset{\boldsymbol{Q}'}{arginf}\ S(\boldsymbol{Q}'), \quad (8)$$

其中，$\boldsymbol{Q}'$ 表示一个一般的 $\boldsymbol{Q}$ 矩阵，其正确性未知，$\hat{\boldsymbol{Q}}$ 表示 $\boldsymbol{Q}$ 矩阵的估计值，“arg inf”表示在整个可能的 $\boldsymbol{Q}$ 矩阵空间中，使 $S(\boldsymbol{Q}')$ 函数取最小值时的 $\boldsymbol{Q}$ 矩阵即为其估计值。

下面介绍适合于前面提到的两种应用情境的算法。

3.2 基于 S_P 统计量的联合估计算法 JE

记测验真实的 $\boldsymbol{Q}$ 矩阵为 $\boldsymbol{Q}_T$，从专家界定的 $\boldsymbol{Q}$ 矩阵(初始 $\boldsymbol{Q}$ 矩阵，其中包含错误) $\boldsymbol{Q}_0$ 出发，即将 $\boldsymbol{Q}_0$ 作为输入，通过联合估计算法，得到 $\boldsymbol{Q}$ 矩阵的估计值 $\hat{\boldsymbol{Q}}$，项目参数 $\hat{\psi}$ 和被试参数 $\hat{\alpha}$，比较 $\hat{\boldsymbol{Q}}$ 与 $\boldsymbol{Q}_T$ 之间的差异，如果 $\hat{\boldsymbol{Q}}$ 与 $\boldsymbol{Q}_T$ 完全一致，则表明算法成功估计，并且计算各参数的估计精度；否则估计不成功。联合估计算法具体过程如下所示：

(1)基于 $\boldsymbol{Q}_0$，作答数据 X，分别采用 EM，MAP 算法估计项目参数和被试参数(Chen & de la Torre, 2013)，并计算 $S(\boldsymbol{Q}_0)$。

(2)在 $\boldsymbol{Q}_0$ 中，固定其它项目，对项目 j，取其属性向量为 q_j' (可能的取值空间记为 Ω^q，有 M^K-1 种取值)，得到 $\boldsymbol{Q}_0^j$，估计参数，并计算 $S(\boldsymbol{Q}_0^j)$，取 $S(\boldsymbol{Q}_0^j)$ 最小时对应的 q_j' 作为项目 j 的属性向量，即：

$$\hat{q}_j = \underset{q_j\subset\Omega^q}{\arg min}(S(\boldsymbol{Q}_0^j))$$

(3)当所有项目都完成估计，记为一次迭代，此时得到的 $\boldsymbol{Q}$ 矩阵记为 $\boldsymbol{Q}(m)$，如果 $\boldsymbol{Q}(m)$ 与 $\boldsymbol{Q}_0$ 完全一致，转到步骤(5)；否则迭代次数加 1，转到步骤(4)。

(4)将 $\boldsymbol{Q}(m)\to\boldsymbol{Q}_0$，重复步骤(2)。

(5)算法结束，输出 $\boldsymbol{Q}(m)$ 和此时的参数估计值 $\hat{\psi}$ 和 $\hat{\alpha}$。

3.3 基于 S_P 统计量的在线估计算法 OE

JE 算法需要专家已经对测验中的所有项目属性均已界定，只是其中包含错误。不同的是，OE 算法只需要专家对少部分项目已经界定，对剩余的项目未界定(可以是以下三种情况：新编制的项目需要界定属性、专家之间对属性界定持不同意见的项目、属性定义不确定或有怀疑的项目)，在这种情况下，可以采用 OE 算法进行估计。

chinaXiv:202303.08463v1

ChinaXiv合作期刊

记已界定属性向量的这部分项目为 Q_0^{base}，剩余需要界定的项目记为 Q_0^{New}，因为 Q_0^{base} 部分已经界定，每次从 Q_0^{New} 中无放回地取出一个项目(记为 q_0)加入到 Q_0^{base} 中，估计 q_0 的属性向量 q_0，将 $\hat{q}_0 \bigcup Q_0^{base} \to Q_0^{base}$；重复这个过程，直到 Q_0^{new} 为空。在线估计算法的具体过程如下所示：

(1)从 Q_0^{new} 中无放回地取出一个项目加入到 Q_0^{base} 中，为方便介绍且不失一般性，假设新加入的项目总是放在第一行，即 $q_0 \cup Q_0^{base} \to Q_0^{base}$。

(2)基于 Q_0^{base}，作答数据，估计项目参数和被试参数，并计算 $S(Q_0^{base})$。

(3)在 Q_0^{base} 中，对新加入的项目，取其属性向量为 q_j'，估计参数，并计算 $S(Q_0^{base})$，取 $S(Q_0^{base})$ 最小时对应的 q_j' 作为项目 j 的属性向量，即：

$$\hat{q}_j = \underset{q_j \subset \Omega^q}{\arg min}(S(Q_0^{base}))$$

(4)如果 Q_0^{new} 不为空，重复步骤(1)；否则转到步骤(5)。

(5)算法结束，输出 Q_0^{new} 的估计值 $\hat{Q}_0^{new}$。

需要说明的是，当初始时 Q_0^{base} 完全正确且包含适当数量时，这种“增量式”的 OE 算法每次只对第一个项目进行估计，可以有效地避免了一次加入多个错误项目所带来的“遮罩效应(masking effect; Fung, 1993; Yuan & Zhong, 2008)”；但是如果 Q_0^{base} 也包含错误或者数量较少时，OE 算法对部分项目的估计仍可能包含错误，此时需要对 OE 算法的输出 Q 矩阵采用 JE 算法进行估计，即采用“二次校正”的方法来保证估计的正确性。

4　研究设计

为了评价基于 S 统计量的两种算法对于多级属性 Q 矩阵估计的表现，我们通过模拟研究考察它们在不同的条件下成功识别正确 Q 矩阵的可能性。如前面所述，假设手头已有专家界定的 Q 矩阵 Q_0，这个 Q 矩阵中可能存在少量的错误，为考察 S_P 统计量的表现，分两种情况，第一种情况：Q_0 中的属性向量被界定错误仅仅是部分属性的值存在大小上的错误，即将属性的值过高的设定或过低(但不包括 0)的设定，比如 q_{jk} 的值应该为 2，但实际上专家将其界定为 1 或 3；第二种错误情况：既存在属性的值在大小上的错误，也存在误设未考察的属性或缺失考察的属性，比如：误将 $q_j = (2\,1\,0\,0\,0)'$ 设置为 $(1\,0\,0\,1\,0)'$。在不引起误解的情况下，下文将这两种类型的错误分别简称为错误 I 和错误 II，将错误 I 和错误 II 所对应的可能的属性向量空间分别记为 Ω_1和Ω_2。可以看出，错误 II 是实际测验中的一般情形，错误 I 是它的一种特殊的情形。

由于本研究中被试可能的属性掌握模式数为 $3^5 = 243$，如果被试人数为 500 的均匀分布总体，则平均每类被试数量偏少，仅为 2.06，故样本量最小取 1000 人。

4.1　对于 JE 算法

为了研究 S_P 统计量在多级属性 Q 矩阵估计中的表现，考虑的因素有：项目个数，测验人数，包含错误的项目个数和错误项目的类型共四个因素，其中项目个数参考 Chen 和 de la Torre (2013)关于多值属性 Q 矩阵的设定，包括 2 个水平，分别为 15 和 30，测试人数(3 个水平，1000, 2000 和 4000)，包含错误的项目类型(2 个水平，错误 I 和错误 II)和错误的项目个数。错误的项目个数参考 Liu 等人(2012)的设置，共 3 个水平，分别为 3，4 和 5，表明“初始的 Q 矩阵”中包含 3, 4 或 5 个属性向量被错误标定的题目，余下题目的属性向量都是被正确标定的。因此，一共有 $2\times3\times2\times3 = 36$ 个实验条件。

4.2　对于 OE 算法

JE 算法中，已假定专家对所有的 J 个项目都进行了属性向量界定，OE 与 JE 算法不一样的是专家只是对“基础项目”部分进行了界定，余下的是需要估计的“新项目”。本研究中这部分“新项目”的属性向量初值是随机生成的。这里考察的因素主要有：项目个数(与 JE 算法一样，仍然是 2 个水平，分别为 15，30)，测试人数(与 JE 算法相同，还是 3 个水平，1000, 2000 和 4000)，基础项目个数参考 Qin 等人(2015，2020)的设置，其中测验长度为 30 时有 8 个水平，分别是 8, 9, 10, 11, 12, 13, 14 和 15；测验长度为 15 时有 6 个水平，分别是 5, 6, 7, 8, 9, 10。因此，一共有 $3\times8+3\times6 = 42$ 个实验条件。

4.3　数据模拟

4.3.1　Q 矩阵

测验的真实 Q 矩阵如网络版附录中的表 A1 和表 A2 所示，Q_1 中共有 30 个项目，Q_2 中有 15 个项目(Chen & de la Torre, 2013; Yu & Cheng, 2020)。为方便区分，将包含 30，15 个项目的 Q 矩阵记为 Q_1^{30}, Q_2^{15}。

4.3.2　项目参数

项目参数假设服从均匀分布，猜测参数和失误参数都按 $U(0.05,0.20)$ 模拟。

chinaXiv:202303.08463v1

ChinaXiv合作期刊

chinaXiv:202303.08463v1

4.3.3 被试参数

被试的知识状态分布按均匀分布模拟，即 M^K 种知识状态的被试人数相近。

4.3.4 作答数据

基于真实的 $\boldsymbol{Q}$ 矩阵、项目参数和被试参数，按照 p-DINA 模型模拟作答数据。

4.3.5 初始 $\boldsymbol{Q}$ 矩阵

(1)对于 JE 算法，随机从真实 $\boldsymbol{Q}$ 矩阵的 J 个项目中选出若干个项目，并将其属性向量按照预定方案(错误 I 和错误 II)修改成错误的状态，但不能是全 0 的向量，也不能是其正确的值，将修改后的矩阵作为“初始 $\boldsymbol{Q}$ 矩阵”。

(2)对于 OE 算法，随机从真实 $\boldsymbol{Q}$ 矩阵的 J 个项目中选出若干个项目作为“基础项目”，而余下的项目作为“新”项目，“新”项目的属性向量初始值按随机方式生成，但不能是 0 向量，也不能是其正确的向量。

4.3.6 参数估计

数据的模拟和分析采用 matlab 编写程序完成，每种实验条件重复 100 次，最后取 100 次的平均值作为最终的结果。

4.3.7 评价指标

这里采用三个指标来评价多级属性 $\boldsymbol{Q}$ 矩阵估计算法的表现，分别是：$\boldsymbol{Q}$ 矩阵成功恢复率、平均迭代次数和平均执行时间。$\boldsymbol{Q}$ 矩阵成功恢复率是指在某种条件下的 100 批数据中，算法输出的 $\boldsymbol{Q}$ 矩阵完全匹配真实 $\boldsymbol{Q}$ 矩阵的比率，计算公式为

$$r_Q = \frac{\sum_{t=1}^{T} I(\hat{\boldsymbol{Q}}, \boldsymbol{Q}_T)}{T}, \tag{9}$$

这里 T 为实验重复次数，这里取 100，I 为示性函数，$I(\hat{\boldsymbol{Q}}, \boldsymbol{Q}_T)$ 在 $\boldsymbol{Q}$ 和 $\boldsymbol{Q}_T$ 完全一致时取 1，否则取 0。

平均迭代次数是对 100 次估计的总迭代次数计算平均值。

$$ANI_Q = \frac{\sum_{t=1}^{T} ite_t}{T}, \tag{10}$$

这里 ite_t 表示第 t 批数据需要的迭代次数。

与平均迭代次数类似，我们同样也分别记录了两种方法的平均执行时间，它也描述了对应方法的计算效率，具体计算公式为

$$ART_Q = \frac{\sum_{t=1}^{T} time_t}{T}, \tag{11}$$

这里 $time_t$ 表示第 t 批数据需要的执行时间，以秒为单位。上面的三个指标中，r_Q 描述的是算法的估计精度，值越大表示算法的估计精度越高。ANI_Q 和 ART_Q 描述的是算法的运行效率，值越小表明算法的效率越高。

4.4 研究 1：多级属性 $\boldsymbol{Q}$ 矩阵和参数的联合估计

联合估计适合的测验情形是：专家已对测验项目都已界定，只是对部分项目的属性定义尚不确定、可能界定错误或意见不统一时使用。采用 JE 算法可以对 $\boldsymbol{Q}$ 矩阵进行验证，并且输出建议的 $\boldsymbol{Q}$ 矩阵。下面分两种错误类型进行介绍。

4.4.1 仅仅存在属性值界定错误时的联合估计

在实际应用中，专家在界定某些项目的属性值时出现分歧或错误的情况，即前面所介绍的错误 I，这是一种相对简单的情形。因此本研究考察当初始 $\boldsymbol{Q}$ 矩阵中有部分项目仅仅出现属性低估或高估的情况(不包括低估至 0 或从 0 高估的情况)。

学生在测验中的作答模拟是按真实 $\boldsymbol{Q}$ 矩阵完成，只是在分析数据时采用包含错误的“初始 $\boldsymbol{Q}$ 矩阵”作为输入，采用 JE 算法来实现对 $\boldsymbol{Q}$ 矩阵、项目参数和被试参数的联合估计，最后比较算法估计得到的 $\boldsymbol{Q}$ 矩阵与真实 $\boldsymbol{Q}$ 矩阵之间的差异，若完全一致，则估计成功，否则估计失败，并且统计估计过程中的迭代次数。

4.4.2 存在属性值错误、含多余属性或缺失必要属性时的联合估计

相对来说，错误 II 是比错误 I 更严重的错误，因为不但会出现属性低估和高估，同时还会出现将未考察的属性包含进来，也可能会出现将考察的属性遗漏，这在实际应用也是有可能出现的，错误 I 可以看成是错误 II 的一种特殊情形。因此本研究考察当初始 $\boldsymbol{Q}$ 矩阵出现错误 II 时 JE 算法的表现。

4.5 研究 2：多级属性 $\boldsymbol{Q}$ 矩阵和参数的在线估计

在线估计算法 OE 适合的另一种测验情形，即仅仅少部分项目被正确界定，有大批项目需要定义属性向量的情况，比如对编制的一批新题进行界定(包括属性向量和参数)，“新项目”的属性向量不需要专家进行初始界定，可以按随机方式生成，在这种情况下，可以借助已有项目的信息，完成对新项目的界定。

界定时需要学生同时作答“基础项目”和“新项目”，估计时固定“基础项目”的属性向量，只需要估计“新项目”的属性向量。为了充分利用已有信息，减少“噪音”信息引起的“遮罩效应”(masking effect;

ChinaXiv合作期刊

Fung, 1993; Yuan & Zhong, 2008)带来的负面影响，估计时采用每次只加入一个“新项目”的增量式估计的方式进行。并且，为了降低由于“基础题”的质量所带来的影响，在 OE 算法结束后，对整个 ***Q*** 矩阵再使用 JE 算法进行整体估计，提高估计的成功率。最后比较算法估计得到的 ***Q*** 矩阵与真实 ***Q*** 矩阵之间的差异，若完全一致，则估计成功，否则估计失败，并且统计估计过程中的迭代次数。

需要注意的是，OE 算法中是指完成所有的“新项目”估计后，如果“新项目”没有估计成功，则对包含“基础项目”和“新”项目的 ***Q*** 矩阵用 JE 算法进行联合估计，因此从这个角度来看，OE 算法中的迭代次数与 JE 算法中一样，也是指对所有项目完成一次估计的次数。

4.6　试验结果

4.6.1　JE 算法的估计结果

表 1~表 4 是 JE 算法在项目数(30, 15)和错误类型(I 和 II)时的估计结果，从结果可以看出，JE 算法在估计 ***Q*** 矩阵时，其执行效率和正确率受到多方面因素的影响，比如：被试人数，测验的项目数，包含的错误项目数等的影响。研究 1 和研究 2 是分别安排在两台云服务器上运行的，服务器的具体配置是：CPU 是 2 颗至强 E5-2697，十二核心；内存类型 DDR5，容量是 64 G；硬盘类型是固态，容量 512 G。从算法的执行效率来看，虽然算法的搜索空间已经下降了很多，但是依然有较大的搜索空间，各种条件下的平均执行时间仍然较大，最低情况下需要一天的时间(89182.33 秒)。从算法的正确率来看，相对来看，测验项目数对于正确率的影响很大，测验项目从 30 下降到 15，估计成功率平均下降了 61.67%。

从表 1 和表 2 中可以看出，被试人数和测验项目数都与 ***Q*** 矩阵估计成功率有正向的相关关系，而错误项目数与 ***Q*** 矩阵估计成功率则有负向的相关关系。根据本研究中的条件，被试人数为 2000，测验项目数为 30，可以达到较好的估计结果。具体来说，对于估计成功率，***Q*** 矩阵包含 30 题时各条件下都能达到 80%以上，而 15 题时最好的情况都要小于 60%。从迭代次数来看，测验项目数为 15 时，各样本条件下需要的平均迭代次数小于 2.5，而当项目数达到 30 时，对应需要的迭代次数超过 3。图 1 和图 2 进一步展示了 JE 算法的表现随着错误界定项目数发生变化的趋势。

表 3 和表 4 分别是测验项目数为 30，15，并且 ***Q*** 矩阵中包含错误类型 II 时的估计结果。可以看出，一方面被试人数的增加可以提高 JE 算法的估计成功率，比如测验长为 30，错误项目数为 3 和 5 时，被试人数从 1000 提高到 4000，估计成功率分别提高了 7%和 13%。另一方面，被试人数和错误项目数会对估计成功率会产生交互作用，因为当测验长度只有 15，错误项目数 3 和 5，人数从 1000 提高到 4000，估计成功率分别提高了 18%和 5%，此时人

表 1　错误类型 I，Q_1^{30} 时 JE 算法的估计成功率和平均迭代次数

包含的错误项目数	被试人数								
	1000			2000			4000		
	成功率(%)	平均迭代次数	平均执行时间(s)	成功率(%)	平均迭代次数	平均执行时间(s)	成功率(%)	平均迭代次数	平均执行时间(s)
3	94	2.05	197397.50	98	2.04	205128.46	98	2.00	211650.67
4	92	2.14	210386.75	95	2.12	208827.46	96	2.14	213588.22
5	81	2.30	234271.81	94	2.19	211649.67	94	2.21	215590.22

表 2　错误类型 I，Q_1^{15} 时 JE 算法的估计估计成功率和平均迭代次数

包含的错误项目数	被试人数								
	1000			2000			4000		
	成功率(%)	平均迭代次数	平均执行时间(s)	成功率(%)	平均迭代次数	平均执行时间(s)	成功率(%)	平均迭代次数	平均执行时间(s)
3	36	3.13	89182.33	46	3.02	101401.32	54	2.92	109542.61
4	21	3.63	90511.47	27	3.44	111399.52	38	3.33	115674.36
5	18	3.89	135365.82	22	3.62	138115.65	25	3.47	144921.76

chinaXiv:202303.08463v1

ChinaXiv合作期刊

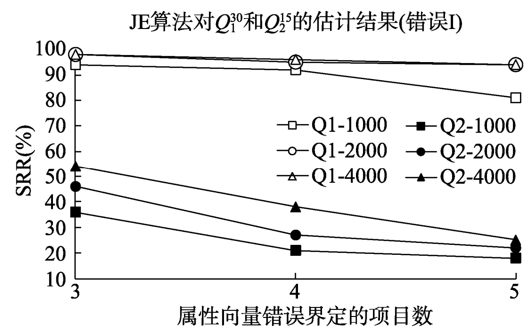


图 1 错误类型 I 时, JE 算法的估计结果

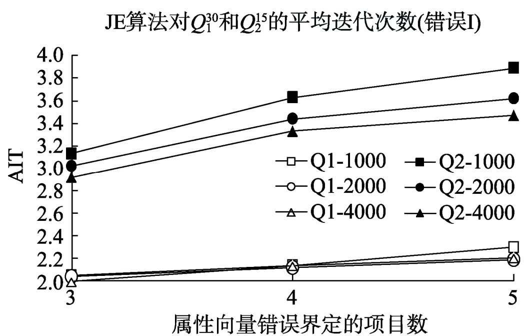


图 2 错误类型 I 时, JE 算法的迭代次数

表 3 错误类型 II, Q_1^{30} 时 JE 算法的估计成功率平均迭代次数

包含的错误项目数	被试人数								
	1000			2000			4000		
	成功率(%)	平均迭代次数	平均执行时间(s)	成功率(%)	平均迭代次数	平均执行时间(s)	成功率(%)	平均迭代次数	平均执行时间(s)
3	91	2.94	217999.39	97	2.38	207354.22	98	2.43	212017.28
4	90	3.17	221085.75	95	2.58	209615.68	96	2.68	214643.81
5	80	3.75	254841.29	89	3.32	242336.01	93	3.58	287900.52

表 4 错误类型 II, Q_2^{15} 时 JE 算法的估计成功率和平均迭代次数

包含的错误项目数	被试人数								
	1000			2000			4000		
	成功率(%)	平均迭代次数	平均执行时间(s)	成功率(%)	平均迭代次数	平均执行时间(s)	成功率(%)	平均迭代次数	平均执行时间(s)
3	33	3.34	92723.60	45	3.32	101737.07	51	3.28	119922.70
4	17	3.85	97788.49	25	3.74	111740.98	37	3.73	126056.07
5	15	4.41	144782.21	18	4.32	184428.18	20	4.27	195388.36

数的增加对低错误项目数影响更大, 这与测验长度为 30 时的情况正好相反。图 3 和图 4 是测验项目为 15 题时 JE 算法的表现随着错误界定项目数发生变化的情况。

综合表 1, 表 2, 表 3 和表 4 可以看出, 一方面, 当错误类型为 II 时, 相同人数、题目条件下要略低于错误类型 I 时的估计成功率, 并且相应的迭代次数也要更多, 这是因为错误类型 II 时, 项目属性向量可能的取值空间更大所导致的; 另一方面, 从平均运行时间来看, 相对于错误类型 I, 固定其它条件时错误类型 II 各对应的实验条件需要相对更多

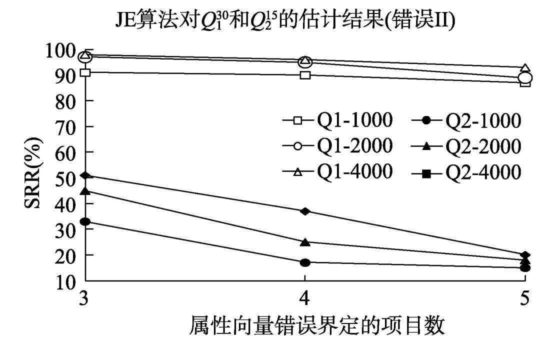


图 3 错误类型 II 时, JE 算法的估计结果

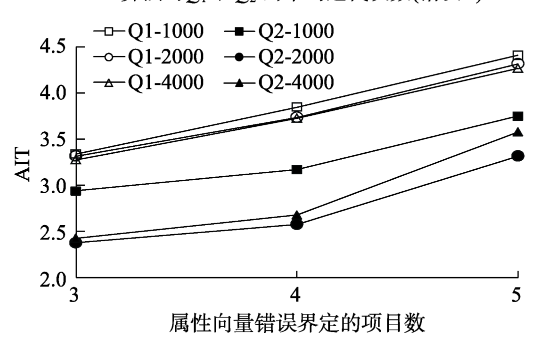


图 4 错误类型 II 时, JE 算法的迭代次数

chinaXiv:202303.08463v1

ChinaXiv合作期刊

的运行时间, 这一点是和更大的迭代次数相一致的。

综合图 1、图 2、图 3 和图 4, 随着 **Q** 矩阵中包含的错误项目数增加, 不论是错误类型 I 还是错误类型 II, JE 算法估计的成功率在下降, 所需要的迭代次数在增加。

4.6.2　OE 算法的估计结果

表 5 和表 6 分别是 OE 算法在 $\boldsymbol{Q}_1^{30}$ 和 $\boldsymbol{Q}_2^{15}$, 不同基础题时的在线估计结果, 从结果来看, 要想达到较好的估计成功率, 不同被试人数需要的基础项目个数不同, 被试人数越多时需要的基础项目越少, 比如对于 $\boldsymbol{Q}_1^{30}$, 要想达到 90%以上的估计成功率, 1000 人需要 10 个基础题, 2000 人和 4000 人只需要 8 个基础题即可; 要想达到 95 以上的成功率, 1000 人和 2000 人都至少需要 13 个基础题, 而 4000 人只需要 12 个基础题。而对于 $\boldsymbol{Q}_2^{15}$, 要想达到 80%以上的估计成功率, 三种被试人数都需要至少 9 个基础题。对于相同的基础题数, OE 算法对 30 题的 **Q** 矩阵估计的成功率要高于 15 题的 **Q** 矩阵, 这主要是因为题目数增加提高了被试的属性掌握模式估计准确率导致的。当基础题为 10 时, 从图 5 和图 6 来看, 估计成功率是随着基础题的增加而增加, 所需要的迭代次数是随着基础题的增加而减少的, 图 7 和图 8 也显示了相同的变化趋势。从 OE 算法的运行效率来看, 随着"基础项目"的增加, 所需要的运行时间在下降, 比如在测验长为 30, 8 和 15 个"基础项目", 1000 人时, 平均运行时间分别是 176481.88 和 23545.31 秒, 这是因为 OE 算法所花费的时间主要是由"新题"的数量和联合估计决定的, 而联合估计过程的耗时占用了时间的大部分, 8 和 15 个"基础项目"条件下的平均迭代次数分别为 1.78 和 0.22。

从图 5~图 8 可以看出, 当测验项目数从 30 降到 15 时, 算法所需要的迭代次数会有较大的增加, 比如基础题为 10 个, 1000 人, 长度 30 和 15 的测验所需要的迭代次数分别为 0.74 和 1.06。

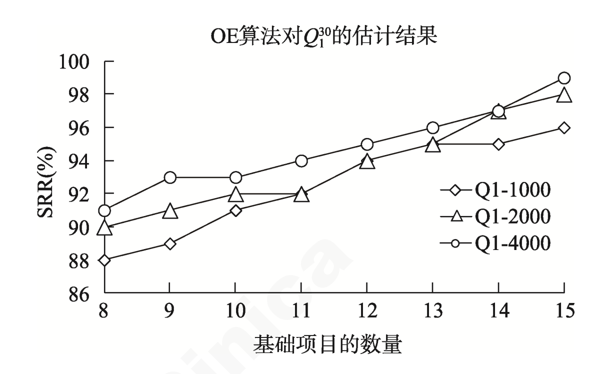


图 5　OE 算法在 $\boldsymbol{Q}_1^{30}$ 的估计结果

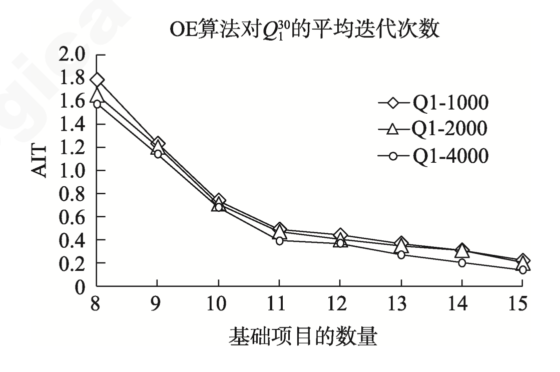


图 6　OE 算法在 $\boldsymbol{Q}_2^{15}$ 的迭代次数

表 5　$\boldsymbol{Q}_1^{30}$ 时 OE 算法的估计成功率和平均迭代次数

包含的基础项目数	被试人数 1000			2000			4000		
	成功率(%)	平均迭代次数	平均执行时间(s)	成功率(%)	平均迭代次数	平均执行时间(s)	成功率(%)	平均迭代次数	平均执行时间(s)
8	88	1.78	176481.88	90	1.65	166386.66	91	1.57	171756.48
9	89	1.23	118728.54	91	1.20	123921.95	93	1.14	122017.14
10	91	0.74	72991.02	92	0.71	78193.55	93	0.68	74526.40
11	92	0.49	49849.71	92	0.47	51103.55	94	0.39	41299.84
12	94	0.44	43077.71	94	0.40	45441.27	95	0.37	42427.26
13	95	0.37	38305.11	95	0.35	40129.54	96	0.27	30554.28
14	95	0.31	31613.60	97	0.31	33325.60	97	0.20	22460.50
15	96	0.22	23545.31	98	0.20	24116.44	99	0.14	15503.14

注：OE 算法中的平均迭代次数是指在对数据进行整体估计时的平均迭代次数, 如果估计过程不需要整体估计即已成功完成, 则该批数据的迭代次数为 0。

chinaXiv:202303.08463v1

ChinaXiv合作期刊

表 6 Q_2^{15} 时 OE 算法的估计估计成功率和平均迭代次数

包含的基础项目数	被试人数								
	1000			2000			4000		
	成功率(%)	平均迭代次数	平均执行时间(s)	成功率(%)	平均迭代次数	平均执行时间(s)	成功率(%)	平均迭代次数	平均执行时间(s)
5	37	1.98	59247.69	46	1.65	60889.27	57	1.57	65979.07
6	45	1.73	51665.79	61	1.50	59236.04	63	1.44	58697.26
7	56	1.54	51053.47	69	1.47	54194.22	72	1.39	52665.52
8	74	1.59	52259.96	77	1.41	47412.48	79	1.38	57552.01
9	81	1.24	37851.94	85	1.14	42252.64	91	1.07	42516.31
10	89	1.06	30857.39	91	1.05	37500.04	93	1.01	40903.18

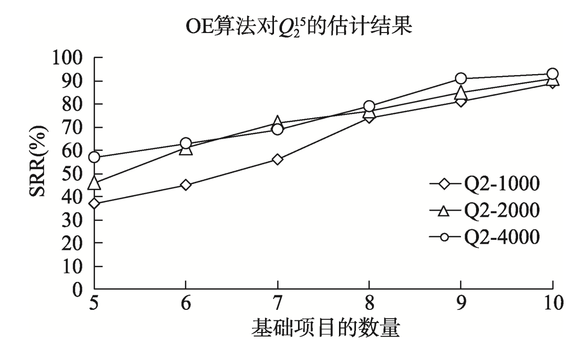


图 7 OE 算法对 $\boldsymbol{Q}_2$ 的估计结果

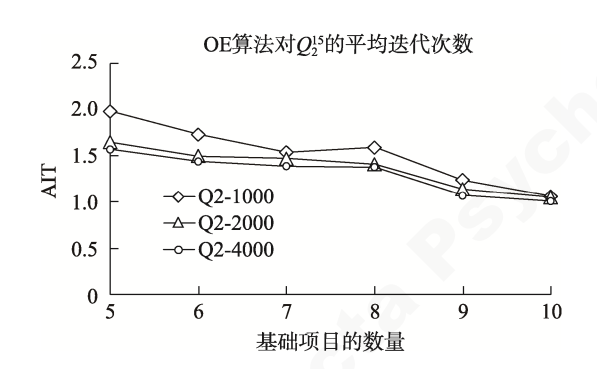


图 8 OE 算法对 $\boldsymbol{Q}_2$ 的平均迭代次数

5 实证数据分析

为了进一步评价两种算法的性能，将它们应用到一批实证数据上。这批实证数据是来自于某市高中的一次月考，选取了数学试卷中与概率有关的试题。这部分测试题考察了随机事件，样本空间，古典概率，使用频数估计概率共 4 个属性。每个属性有 5 个连续的掌握类别：不了解，了解，理解，掌握和应用，分别用 0, 1, 2, 3, 4 表示。基于这 4 个属性，由学科专家共编制了 20 个题，一共有 1960 个考生完成了测验。

以专家界定的“初始 $\boldsymbol{Q}$ 矩阵”(表 7)作为输入，分别采用前面提出的两个算法来验证或估计 $\boldsymbol{Q}$ 矩阵。对于 JE 算法，终止时总的迭代的次数为 4，这个结果比前面模拟研究中的迭代次数要多，这也表明对于实际的测验数据通常是需要更多次的迭代才能够达到算法的收敛条件。JE 算法估计得到的“建议 $\boldsymbol{Q}$ 矩阵”如网络版附录中的表 A3 所示。可以看出，一方面，JE 算法建议修改 6 个题目，共涉及

表 7 概率数据对应的原始 Q 矩阵

题目编号	属性 1	属性 2	属性 3	属性 4	题目编号	属性 1	属性 2	属性 3	属性 4
1	1	1	0	0	11	0	0	4	2
2	0	0	2	1	12	0	4	0	1
3	0	3	0	4	13	2	0	3	0
4	0	0	2	0	14	0	1	0	3
5	1	2	0	0	15	2	1	0	0
6	0	1	1	0	16	0	1	1	0
7	0	2	0	0	17	0	2	0	0
8	3	0	0	1	18	4	0	0	1
9	1	1	0	0	19	0	0	4	2
10	1	2	0	0	20	0	1	0	1

chinaXiv:202303.08463v1

ChinaXiv合作期刊

到 7 个属性, 并且对这 7 个属性都是属性水平上的修改, 即认定初始 ***Q*** 矩阵出现了错误类型 I。另一方面, 参数估计的结果表明考生的属性掌握模式不是均匀分布的, 整个数据中只出现了 76 种属性掌握模式。

对于 OE 算法, 我们选择了初始 ***Q*** 矩阵中的 5 个题目(表 A4 中灰色背景显示的题目), 选择这 5 个题目的原因是学科专家对这 5 道题的属性定义完全一致, 并且它们在 JE 算法的建议 ***Q*** 矩阵中也得到了验证。余下的 15 道题作为"新题目", 将它们逐个用 OE 算法进行估计。当所有的"新题目"完成了估计, 再用JE算法对所有题目进行联合估计, 这样就得到了 OE 算法建议的 ***Q*** 矩阵, 如网络版附录中的表 A4 所示。可以看出, OE 算法建议修改 6 个题目, 共涉及 6 个属性。除了第 19 题之外, 由 JE 和 OE 两种算法得到的建议 ***Q*** 矩阵是完全一致的。对于第 19 题, 专家界定的初始向量为[0 0 4 2], JE 和 OE 算法得到的属性向量分别是[0 0 3 3]和[0 0 4 3]。在与 5 位一线的教师进行讨论之后, 他们其中的 4 位都倾向于同意 OE 算法得到的结果, 即将第 4 个属性初始定义的水平 2 修改为水平 3。

6 讨论与进一步的研究方向

本研究将适合二级属性 ***Q*** 矩阵的 *S* 估计量拓展到多级属性的 ***Q*** 矩阵估计中, 使得多级属性 ***Q*** 矩阵的验证和估计成为可能, 并针对实际应用中的两种常见情境, 分别介绍了两种算法: 即 JE 和 OE 算法, 它们分别适用于不同的场合。当手头已有 ***Q*** 矩阵的初值(可以由专家来初步界定)时, 可以采用 JE 算法进行验证, 而 OE 算法是当手头只有少部分项目的属性向量已经界定, 需要对更多的项目进行定义时使用。模拟实验结果表明, 尽管多级属性 ***Q*** 矩阵的搜索空间相对于二级属性 ***Q*** 矩阵更大, 但这两种算法在各自适用的情况下都有较高的估计成功率。

虽然 JE 和 OE 算法在模拟条件下取得了较好的结果, 即使如此, JE 和 OE 算法仍然需要在更复杂的情况中去验证, 对于 JE 算法, 这里只考虑"初始 ***Q*** 矩阵"中包含的错误项目较少, 对于更多错误时的估计或者所能容忍的最大错误项目数量需要进一步研究; 对于 OE 算法, 研究中随机选择了 100 批"基础项目", 这 100 批"基础项目"的质量有好有坏, 并没有考虑"基础项目"的质量对于估计的影响, 如果进一步研究"基础项目"的设计, 使之更有利于"新项目"的估计, 就像诊断测验中的 ***Q*** 矩阵设计一样, 在基础题中加入"可达矩阵"对于 ***Q*** 矩阵估计的影响等(Chen et al., 2015; 丁树良 等, 2019; 彭亚风 等, 2016, 2018; Gu et al., 2018; Gu & Xu, 2021), 应该是很有意义的工作。本研究中无论是 JE 还是 OE 算法, 只考虑了两种错误类型, 实际上, 还有可能存在其它的错误类型, 未来需要对其它更多可能的情况进行研究。另外, 现实的测验情境往往是很复杂的, 比如考生可能是存在多种解题策略的, 因此, 结合多种策略的诊断测验中 ***Q*** 矩阵的估计需要进一步考虑(黄玉 等, 2019)。测验的属性间很可能存在某种层级关系(喻晓锋 等, 2021), 属性间存在层级关系时的多值 ***Q*** 矩阵估计也是未来需要研究的方向。

基于 S 统计量的 ***Q*** 矩阵估计一个不足之处在于需要花费较多的时间, 这对于实际应用可能是一个潜在的缺陷, 未来对提出的方法进行时间效率上的改进或研究时间效率更高的方法都值得进一步研究。比如 Yu 和 Cheng (2020)的研究表明, 0-1 计分下基于残差统计量的统计量比基于 *S* 统计量在运行效率上有优势, 因此将基于残差的统计量拓广到多值属性诊断测验的 ***Q*** 矩阵估计值得考虑; 未来也需要进一步考虑一些非参数的方法, 因为它们通常对于样本量的要求较小, 并且有执行效率上的优势(刘娜 等, 2021); 将基于深度学习等一些算法拓广到多值属性诊断测验的 ***Q*** 矩阵估计(张玉柳 等, 2021; Li et al., 2022)也需要深入研究。

实证数据的分析表明, 本研究中提出的基于 *S* 统计量的联合估计算法和在线估计算法可以在实际中应用, 并且结果显示专家对于题目属性向量的错误定义更容易出现在高估或低估属性的水平上, 不太容易出现完全缺失某个属性或包含额外的属性等更严重的情况。OE 算法的一个副产品是同时将新项目的参数进行了估计, 并且它能保证与基础项目的参数处于同一个尺度上。将属性间的关系纳入考虑需要进一步研究, 未来也需要将算法应用到其它的诊断模型中(Ma & de la Torre, 2019; Zhan et al., 2020)。

参 考 文 献

Cai, Y., & Tu, D. B. (2015). Extension of cognitive diagnosis models based on the polytomous attributes framework and their Q-matrices designs. *Acta Psychologica Sinica, 47*(10), 1300–1310.

[蔡艳, 涂冬波. (2015). 属性多级化的认知诊断模型拓展及其 Q 矩阵设计. *心理学报, 47*(10), 1300–1310.]

Chen, J. S., & de la Torre, J. (2013). A general cognitive

chinaXiv:202303.08463v1

ChinaXiv合作期刊

diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*(6), 419–437.
Chen, Y. X., Liu, J. C., Xu, G. J., & Ying, Z. L. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association, 110*(510), 850–866.
Chung, M.-T. (2014). *Estimating the Q-matrix for cognitive diagnosis models in a Bayesian framework.* (Unpublished doctoral dissertation), Columbia University, New York.
DeCarlo, L. T. (2012). Recognizing Uncertainty in the Q-Matrix via a Bayesian Extension of the DINA Model. *Applied Psychological Measurement, 36*(6), 447–468.
de La Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*(4), 343–362.
de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130.
de la Torre, J. (2011). The generalized dina model framework. *Psychometrika, 76*(2), 179-199.
de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*(2), 253–273.
Ding, S. L., Luo, F., Wang, W. Y., & Xiong, J. H. (2019). The designing cognitive diagnostic test with dichotomous scoring. *Journal of Jiangxi Normal University (Natural Science), 43*(5), 441–447.
[丁树良, 罗芬, 汪文义, 熊建华. (2019). 0-1 评分认知诊断测验设计. *江西师范大学学报(自然科学版), 43*(5), 441–447.]
Fung, W.-K. (1993). Unmasking outliers and leverage points: A confirmation. *Journal of the American Statistical Association, 88*(422), 515–519.
Gu, Y. Q., Liu, J. C., Xu, G. J., & Ying, Z. L. (2018). Hypothesis testing of the Q-matrix, *Psychometrika, 83*(3), 515–537.
Gu, Y. Q., & Xu, G. J. (2021). Sufficient and Necessary Conditions for the Identifiability of the Q-matrix. *Statistica Sinica, 31*, 449–472.
Haberman, S. J., von Davier, M., & Lee, Y.-H. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions* (ETS Research Report no. RR-08-45). Princeton, NJ: Educational Testing Service.
Huang, Y., Luo, F., Xiong, J. H., Ding, S. L., & Gan, D. W. (2019). The multiple-strategy cognitive diagnosis method with polytomous scoring. *Journal of Jiangxi Normal University (Natural Science), 43*(4), 376–381.
[黄玉, 罗芬, 熊建华, 丁树良, 甘登文. (2019). *多级评分多策略认知诊断方法. 江西师范大学学报(自然科学版), 43*(4), 376–381.]
Karelitz, T. M. (2004). *Ordered category attribute coding framework for cognitive assessments.* (Unpublished doctoral dissertation), University of Illinois at Urbana-Champaign.
Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.
Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*(3), 205–237.
Li, C. C., Ma, C. C., & Xu, G. J. (2022). Learning large Q-matrix by restricted Boltzmann machines. *Psychometrika*. https://doi.org/10.1007/s11336-021-09828-4.
Liu, J. C., Xu, G. J., & Ying, Z. L. (2012). Data driven learning of Q matrix. *Applied Psychological Measurement, 36*(7), 548–564.
Liu, J. C., Xu, G. J., & Ying, Z. L. (2013). Theory of self-learning Q-matrix. *Bernoulli, 19*(5A), 1790–1817.
Liu, N., Liu, X. L., Li, J. J., Zeng, P. F., Yu, X. J., & Kang, C. H. (2021). Constructing a non-parametric Q-matrix correction method based on Manhattan distance. *Journal of Jiangxi Normal University (Natural Science), 45*(6), 634–641.
[刘娜, 刘芯伶, 李俊杰, 曾平飞, 俞向军, 康春花. (2021). 基于曼哈顿距离构建非参数 Q 矩阵修正方法. *江西师范大学学报(自然科学版), 45*(6), 634–641.]
Luo, Z. S. (2019). *Fundamentals of cognitive diagnostic assessment*. Beijing Normal University publishing group.
罗照盛. (2019). *认知诊断评价理论基础*. 北京师范大学出版集团.
Ma, W., & de la Torre, J. (2019). An empirical Q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology, 73*(1), 142–163.
Peng, Y. F., Luo, Z. S., Li, Y. J., Gao, C. L. (2018). Optimization of test design for examinees with different cognitive structures. *Acta Psychologica Sinica, 50*(1), 130–140.
[彭亚风, 罗照盛, 李喻骏, 高椿雷. (2018). 不同认知结构被试的测验设计模式. *心理学报, 50*(1), 130–140.]
Peng, Y. F., Luo, Z. S., Yu, X. F., Gao, C. L., Li, Y, J. (2016). The optimization of test design in Cognitive Diagnostic Assessment. *Acta Psychologica Sinica, 48*(12), 1600–1611.
[彭亚风, 罗照盛, 喻晓锋, 高椿雷, 李喻骏. (2016). 认知诊断评价中测验结构的优化设计. *心理学报, 48*(12), 1600–1611.]
Qin, C. Y., Jia, S., Fang, X. W., & Yu, X. F. (2020). Relationship validation among items and attributes, *Journal of Statistical Computation and Simulation, 90*(18), 3360–3375
Qin, C. Y., Zhang, L., Qiu, D., Huang, L., Geng, T., Jiang, H., ... Zhou, J. (2015). Model identification and Q-matrix incremental inference in cognitive diagnosis. *Knowledge-Based Systems*, *86*, 66–76.
Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.
Shang, Z. R., Erosheva, E. A., Xu, G. J. (2021). Partial-mastery cognitive diagnosis models. *The Annals of Applied Statistics, 15* (3), 1529 –1555.
Sun, J. N., Xin, T., Zhang, S. M., & de la Torre, J. (2013). A polytomous extension of the generalized distance discriminating method. *Applied Psychological Measurement, 37*(7), 503–521.
Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. Routledge.
Templin, J. L. (2004). *Generalized linear mixed proficiency models for cognitive diagnosis.* (Unpublished doctoral dissertation), University of Illinois at Urbana-Champaign.
Templin, J. L., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*(2), 251–275.
Templin, J. L., Bradshaw, L. (2014). The use and misuse of psychometric models. *Psychometrika, 79* (2), 347–354.
Tu, D. B., & Cai, Y. (2015). The development of CD-CAT with polytomous attributes. *Acta Psychologica Sinica, 47*(11), 1405–1414.
[涂冬波, 蔡艳. (2015). 基于属性多级化的认知诊断计算机化自适应测验设计与实现. *心理学报, 47*(11), 1405–1414.]
von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287–307.
von Davier, M., & Lee, Y.-S. (2019). *Handbook of diagnostic classification models*. Cham: Springer International Publishing.
Wang, D. X., Cai, Y, & Tu, D. B. (2020). Q-matrix estimation methods for cognitive diagnosis models: Based on partial known Q-matrix, *Multivariate Behavioral Research*, 1–13. https://doi.org/10.1080/00273171.2020.1746901.
Xiang, R. (2013). *Nonlinear penalized estimation of true Q-Matrix in cognitive diagnostic models.* (Unpublished doctoral dissertation), Columbia University, New York.
Xu, G.-J. (2013). *Statistical inference for diagnostic classification*

chinaXiv:202303.08463v1

ChinaXiv合作期刊

models. (Unpublished doctoral dissertation), Columbia University, New York.

Yu, X. F., & Cheng, Y. (2020). Data-driven Q-matrix validation using a residual‐based statistic in cognitive diagnostic assessment. *British Journal of Mathematical and Statistical Psychology, 73*(1), 145–179.

Yu, X. F., Luo, Z. S., Gao, C. L., Li, Y. J., Wang, R., & Wang, Y. T. (2015a). An item attribute specification method based on the likelihood D2 statistic. *Acta Psychologica Sinica, 47*(3), 417–426.

[喻晓锋, 罗照盛, 高椿雷, 李喻骏, 王睿, 王钰彤. (2015a). 使用似然比 D2 统计量的题目属性定义方法. *心理学报, 47*(3), 417–426.]

Yu, X. F., Luo, Z. S., Qin, C. Y., Gao, C. L., & Li, Y. J. (2015b). Joint estimation of model parameters and Q-matrix based on response data. *Acta Psychologica Sinica, 47*(2), 273–282.

[喻晓锋, 罗照盛, 秦春影, 高椿雷, 李喻骏. (2015b). 基于作答数据的模型参数和 Q 矩阵联合估计. *心理学报, 47*(2), 273–282.]

Yu, X. F., Ma, Y. F., Luo, Z. S., & Qin, C. Y. (2021). The attribute hierarchical structure learning based on K2 algorithm. *Journal of Jiangxi Normal University (Natural Science), 45*(4), 376–383.

[喻晓锋, 马奕帆, 罗照盛, 秦春影. (2021). 基于 K2 算法的属性层级结构学习研究. *江西师范大学学报(自然科学版), 45*(4), 376–383.]

Yuan, K.-H., & Zhong, X. (2008). Outliers, leverage observations, and influential cases in factor analysis: Using robust procedures to minimize their effect. *Sociological Methodology, 38*(1), 329–368.

Zhan, P. D., Bian, Y. F., Wang, L. J. (2016). Factors affecting the classification accuracy of reparametrized diagnostic classification models for expert-defined polytomous attributes. *Acta Psychologica Sinica, 48*(3), 318–330.

[詹沛达, 边玉芳, 王立君. (2016). 重参数化的多分属性诊断分类模型及其判准率影响因素. *心理学报, 48*(3), 318–330.]

Zhan, P. D., Wang, W., Li, X. M. (2020). A partial mastery, higher-order latent structural model for polytomous attributes in cognitive diagnostic assessments. *Journal of Classification, 37*, 328–351.

Zhang, Y. L., Zhao, B., & Tao, J. H. (2021). The study on students' cognitive state based on fuzzy cognitive diagnostic framework. *Journal of Jiangxi Normal University (Natural Science), 45*(5), 452–459.

[张玉柳, 赵波, 陶金洪. (2021). 基于模糊认知诊断模型的学生认知状态研究. *江西师范大学学报(自然科学版), 45*(5),452–459.]

Validation and estimation of expert-defined *Q*-matrix with polytomous attribute

QIN Chunying[1,2], YU Xiaofeng[1]

([1] *School of Psychology, Jiangxi Normal University, Nanchang, 330022, China*)

([2] *School of Mathematics and Information Science, Nanchang Normal University, Nanchang 330032, China*)

Abstract

Cognitive diagnosis has recently gained prominence in educational assessment, psychiatric evaluation, and many other disciplines. Generally, entries in the *Q*-matrix of traditional cognitive diagnostic tests are binary (two levels, defined as 0 and 1). Polytomous attributes (multi-levels, defined as 0, 1, …), particularly those defined as part of the test development process, can provide additional diagnostic information. Compared to binary attributes, polytomous attributes can not only describe the student's knowledge profile, but can provide more extensive details.

As we all know, *Q*-matrix impacts the accuracy of cognitive diagnostic assessment greatly. Research on the effect of parameter estimation and classification accuracy caused by the error in *Q*-matrix already existed, and it turned out that *Q*-matrix gotten from expert definition or experience was more easily subject to be affected by subjective factors, lead to a misspecified *Q*-matrix. Under this circumstance, it’s urgently needed to find more objective polytomous-attribute *Q*-matrix verification and inference methods.

The present research proposes the verification and estimation of expert-defined polytomous attribute *Q*-matrix based on the polytomous deterministic inputs, noisy, ‘‘and’’ gate (p-DINA) model. We intend to extend the methods adapted to binary *Q*-matrix verification and estimation to polytomous attribute *Q*-matrix, and the proposed methods which can be used in different conditions are joint estimation and online estimation. Simulation results show that: the joint estimation algorithm can be applied to the *Q*-matrix validation which needs an initial *Q*-matrix defined by experts, the online estimation algorithm can be applied to online estimate the “new items” based on a certain number of “based items”. Under the various settings in the simulations, the two estimation algorithms can recover the correct polytomous-attribute *Q*-matrix at a high probability. Empirical study also indicates that the two proposed algorithms can be applied in *Q*-matrix validation or estimation for CDA with polytomous attributes.

Keywords　polytomous attribute, *Q*-matrix, p-DINA model, S statistics

chinaXiv:202303.08463v1

ChinaXiv合作期刊

附录：

附表 A1　30 题对应的 Q 矩阵 Q_1^{30}

项目编号	属性				
	属性 1	属性 2	属性 3	属性 4	属性 5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	2	0	0	0	0
7	0	2	0	0	0
8	0	0	2	0	0
9	0	0	0	2	0
10	0	0	0	0	2
11	1	2	0	0	0
12[a]	0	1	2	0	0
13	0	0	1	2	0
14	0	0	0	1	2
15	2	0	0	0	1
16	1	1	0	0	0
17	0	1	1	0	0
18	0	0	1	1	0
19	0	0	0	1	1
20	1	0	0	0	1
21	1	0	2	0	0
22	0	1	0	2	0
23	0	0	1	0	2
24	2	0	0	1	0
25	0	2	0	0	1
26	2	2	0	0	0
27	0	2	2	0	0
28	0	0	2	2	0
29	0	0	0	2	2
30	2	0	0	0	2

chinaXiv:202303.08463v1

Acta Psychologica Sinica

ChinaXiv合作期刊

附表 A2　15 题对应的 Q 矩阵 Q_2^{15}

项目编号	属性				
	属性 1	属性 2	属性 3	属性 4	属性 5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	1	2	0	0	0
7	0	1	2	0	0
8	0	0	1	2	0
9	0	0	0	1	2
10	2	0	0	0	1
11	2	2	0	1	0
12	2	1	0	0	2
13	1	0	2	2	0
14	0	2	1	0	2
15	0	0	2	2	1

附表 A3　由 JE 算法得到概率论数据的建议 Q-matrix

项目编号	属性			
	属性 1	属性 2	属性 3	属性 4
1	1	1	0	0
2	0	0	2	***2***
3	0	3	0	4
4	0	0	2	0
5	1	***1***	0	0
6	0	***2***	1	0
7	0	2	1	0
8	3	0	0	1
9	1	1	0	0
10	1	2	0	0
11	0	0	4	2
12	0	4	0	1
13	***3***	0	1	0
14	0	1	0	3
15	2	***2***	0	0
16	0	1	1	0
17	0	2	0	0
18	4	0	0	1
19	0	0	***3***	***3***
20	0	2	0	1

注：表格中用粗斜体显示的元素表示 JE 算法所修改后的属性取值

chinaXiv:202303.08463v1

Acta Psychologica Sinica

ChinaXiv合作期刊

附表 A4　由 OE 算法得到概率论数据的建议 *Q*-matrix

项目编号	属性			
	属性 1	属性 2	属性 3	属性 4
1	1	1	0	0
2	0	0	2	***2***
3	0	3	0	4
4	0	0	2	0
5	1	***1***	0	0
6	0	***2***	1	0
7	0	2	1	0
8	3	0	0	1
9	1	1	0	0
10	1	2	0	0
11	0	0	4	2
12	0	4	0	1
13	***3***	0	3	0
14	0	1	0	3
15	2	***2***	0	0
16	0	1	1	0
17	0	2	0	0
18	4	0	0	1
19*	0	0	4	***3***
20	0	1	0	1

注：阴影显示对应的题目表示 OE 算法中的“基础题”，余下的题目对应的是需要估计的“新题”。粗斜体显示元素表示 OE 算法所修改后的属性取值。加星号的题目表示由 OE 算法给出的建议值与 JE 算法给出的建议值不一致的题目。

chinaXiv:202303.08463v1

Acta Psychologica Sinica